Chapter 2
Predicting Newcomer Integration in Online Knowledge Communities by Automated Dialog Analysis

Nicolae Nistor, Mihai Dascălu, Lucia Larise Stavarache, Christian Tarnai and Ștefan Trăușan-Matu

Abstract Using online knowledge communities (OKCs) from the Internet as informal learning environments poses the question how likely these communities will be to integrate learners as new members. Such prediction is the purpose of the current study. Based on the approaches of voices interanimation and polyphony, a natural language processing tool was employed for dialog analysis in integrative versus non-integrative blog-based OKCs. Four dialog dimensions were identified: participants' long-term persistence in the discourse, the community response to their participation, their communicative centrality, and their communicative peripherality. Hierarchical clusters built upon these dimensions reflect socio-cognitive structures including central, regular, and peripheral OKC members. While the socio-cognitive structures did not make a significant difference, integrative OKCs display significantly stronger peripherality, community response, and centrality as compared to non-integrative OKCs.

Keywords Knowledge communities · Newcomer integration · Dialog analysis · Social learning analytics

N. Nistor (✉)
LMU, Faculty of Psychology and Educational Sciences, Munich, Germany
e-mail: nic.nistor@uni-muenchen.de

N. Nistor · C. Tarnai
Universität der Bundeswehr München, Faculty of Human Sciences, Neubiberg, Germany
e-mail: christian.tarnai@unibw.de

N. Nistor
Richard W. Riley College of Education and Leadership, Walden University, Minnesota, USA

M. Dascălu · L.L. Stavarache · Ș. Trăușan-Matu
Faculty of Computer Science, University "Politehnica", Bucharest, Romania
e-mail: mihai.dascalu@cs.pub.ro

L.L. Stavarache
e-mail: larise.stavarache@ro.ibm.com

Ș. Trăușan-Matu
e-mail: stefan.trausan@cs.pub.ro

Y. Li et al. (eds.), *State-of-the-Art and Future Directions of Smart Learning*,
Lecture Notes in Educational Technology DOI 10.1007/978-981-287-868-7_2

2.1 Introduction

Online knowledge communities (OKCs) are frequently regarded in educational research and practice as collaborative environments for informal learning [1, 2]. OKCs display multilayered socio-cognitive structures that comprise central, active/regular, and peripheral participation. Central participants assume more responsibility and perform more difficult tasks than peripheral members; therefore, their identity is that of an expert. After a few decades of mainly qualitative research on communities, researchers are beginning to apply quantitative methods, including the identification of socio-cognitive structures, as shown by Nistor et al. [3], who validated an automated dialog analysis tool, ReaderBench [4]. The automated assessment of OKCs is based on the idea that community discourse is tightly connected with socio-cognitive structures, practice, and learning [2]. Further on, the ReaderBench tool is based on Bakhtin's dialogism [5] and on the polyphonic model of discourse [6]. ReaderBench provides several indicators describing the personal and social dimensions of a collaborative dialog, emphasizing dialog coherence and overall coverage of a given topic. These dimensions are strongly correlated with participants' expertise and critical thinking expressed in online, text-based discussions [3].

Using existing OKCs from the Internet for informal learning poses the question *how likely these communities will be to integrate learners as new, legitimate peripheral members* [2]. The present study aims to answer this research question, thus contributing to understanding and predicting this phenomenon, which will further provide a tool and method to make OKCs "smart" [1, 7].

2.2 Methodology

The study explores the socio-cognitive structures of OKCs that were likely versus unlikely to integrate newcomers (in the following called integrative vs. non-integrative OKCs), following three steps: (1) analyze the community discourse using the ReaderBench tool [3, 4]; (2) cluster the community members based on the resulting discourse characteristics; and (3) compare the clustering results in integrative versus non-integrative virtual communities.

The analysis was conducted on the Internet, in blogger communities publicly available on the blogspot.com and wordpress.com platforms. In a prior study, the researchers had posted a request for survey participation in several blog communities. Two situations emerged: one in which the blog participants responded to the request, and another in which the request was ignored or blocked. Consequently, it was assumed that the former group consisted of integrative (n = 10), the latter of non-integrative (n = 12) OKCs. After these N = 22 blogger communities with a total of 8122 participants were chosen for analysis, the community discourse produced during the entire lifetime of each OKC was downloaded and automatically analyzed. No personal data of the participants were collected.

The ReaderBench tool provides 13 dialog indicators: two overall indicators (number of comments, total collaborative dialog quality), one indicator of the individual contribution to the dialog (individual collaborative dialog quality), five indicators of the social contribution to the dialog (number of initiated discussion threads, length of initiated threads, cumulative interanimation of voices, social collaborative dialog quality, social collaborative dialog quality in initiated threads), and five centrality indicators in the sense of social network analysis (Indegree, Outdegree, Closeness, Eccentricity, and Betweenness).

2.3 Findings

Discourse Analysis. The absolute values of the variables ranged in large limits; hence, they were standardized. Further on, they were strongly correlated with each other; therefore, a principal component analysis was performed. Thus, the number of components was reduced to four factors with eigenvalues greater than 1, which explained 86.16 % of the total variance. The four dimensions resulting after oblimin rotation are based on different sets of the initial variables, as follows. Factor 1R is mainly based on the number of initiated discussion threads and the associated interanimation. As such, Factor 1R is related to the *individual long-term discourse persistence*. Factor 2R is only composed of the average length of initiated threads, thus describing the *community response to* one's *participation* in the collaborative dialog. Factor 3R mainly includes the social network analysis variables Indegree and Betweenness, as well as the social collaborative dialog quality; therefore, it refers to the *individual communicative centrality*. Factor 4R consists of the variables Eccentricity and Closeness; therefore, it describes the *individual communicative peripherality*.

Cluster Analysis. In the second step of the analysis, the three dimensions resulting from the principal component analysis (Anderson-Rubin method) were used as input for a hierarchical cluster analysis according to the Ward method with quadratic Euclidian distances. The optimal separation of clusters was reached for the following four clusters.

Firstly, Clusters 4 and 3 are most visible due to participants' long-term discourse persistence (Factor 1R) and communicative centrality (Factor 3R). Cluster 4 consists of $n = 2$ participants with very high persistence and centrality, low communicative peripherality, and who yield with their interventions relatively strong community response. Cluster 3 consists of $n = 4$ participants with relatively high persistence and centrality, lowest peripherality, and who yield with their interventions the strongest community response. For these reasons, Clusters 4 and 1 reunite the *central OKC members*, from which Cluster 4 represents the OKC *core group*, and Cluster 3 the *opinion leaders* (possibly in a negative sense as well, e.g., "trolls"), who can fundamentally differ from the core group.

Secondly, Cluster 2 consists of $n = 1859$ blog members with moderate discourse persistence (Factor 1R), yielding moderate to strong community response (Factor 2R),

and with moderate centrality (Factor 3R) and moderate, i.e., highest peripherality (Factor 4R). These appear to be the *regular or active OKC members*.

Thirdly and finally, the largest cluster, Cluster 1 ($n = 6257$) reunites the least active OKC members, with very low discourse persistence (Factor 1R), yielding weakest community response (Factor 2R), and with lowest communicative centrality (Factor 3R) and peripherality (Factor 4R). Hence, Cluster 1 can be described as *peripheral OKC members*.

Integrative versus Non-Integrative Blogger Communities. By comparing the extracted data between integrative and non-integrative blog communities, it appears that integrative OKCs are characterized by stronger peripherality ($M = 0.41$, SD = 1.34 for integrative, $M = -0.18$, SD = 0.74 for non-integrative communities, $F(1, 8120) = 642.441$, $p < 0.001$), stronger community response ($M = 0.13$, SD = 1.13 for integrative, $M = -0.06$, SD = 0.93 for non-integrative communities, $F(1, 8120) = 60.626$, $p < 0.001$), and somewhat stronger centrality ($M = 0.04$, SD = 0.71 for integrative, $M = -0.02$, SD = 1.10 for non-integrative communities, $F(1, 8120) = 4.911$, $p < 0.05$). Significant differences between integrative and non-integrative communities could be found neither in terms of long-term discourse persistence, nor in terms of socio-cognitive structure (i.e., percent of central, active and peripheral members, and relationships between these).

2.4 Discussion and Conclusions

In summary, this study lays the ground for the educational application of OKCs as informal learning environments. This requires in turn that the OKCs integrate the learners in their community discourse. This study assumes that the integrativity of an OKC is tightly connected to the community discourse and practice; hence, it can be assessed by discourse analysis, as follows.

In the first step, the polyphony-based [6] tool ReaderBench [3, 4] was employed to analyze the blog-based OKC discourse. From the multitude of provided results, the following ground dimensions were extracted: (1) individual participants' long-term persistence in the discourse, (2) the community response to their participation, (3) their communicative centrality, and (4) their communicative peripherality within the social network. These dimensions result from Bakhtin's polyphony theory [5] and Trăuşan-Matu's analytic approach [6]. They describe the interanimation of voices within a collaborative dialog and appear appropriate for automated discourse analysis.

In the second step, the community members were clustered based on their discourse characteristics. The hierarchical cluster analysis offered a classification including central, active/regular, and peripheral OKC members, which corresponds to the socio-cognitive structures described in the CoP research [2].

In the third and final step, the extracted results were compared between integrative and non-integrative OKCs. While there were no significant differences in terms of socio-cognitive structure, integrative OKCs were associated with

significantly stronger communicative peripherality, community response, and communicative centrality.

For educational practice, the conclusions of this study are straightforward: Existing OKCs from the Internet can be used as informal learning environments, for example, in higher education, applying social learning analytics tool such as ReaderBench to optimize the learning activity and make the OKCs "smart" [1, 7]. Appropriate instructional design should be developed and evaluated in the near future. For OKC research, this study adds empirical evidence for the relationship between community discourse and practice.

However, the result validity may be limited by several conceptual and methodological shortcomings. Although the number of participants was fairly high, there was a relatively small number of OKCs involved in the study. Also, integrativity was assimilated to OKC members' response to relatively simple requests. Upcoming research aims to extend both the samples and the perspective on integrativity by observing the long-term interactions between regular OKC members and visitors.

Acknowledgements This work has been partially funded by the 644187 RAGE H2020-ICT-2014 project, as well as by the Sectorial Operational Program Human Resources Development 2007–2013 of the Romanian Ministry of European Funds according to the Financial Agreements POSDRU/159/1.5/S/134397 and 134398.

References

1. Scardamalia, M., & Bereiter, C. (2014). Smart technology for self-organizing processes. *Smart Learning Environments, 1*, 1–13.
2. Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press
3. Nistor, N., Trăuşan-Matu, Ş., Dascălu, M., Duttweiler, H., Chiru, C., Baltes, B., & Smeaton, G. (2015). Finding student-centered open learning environments on the internet: Automated dialogue assessment in academic virtual communities of practice. *Computers in Human Behavior, 47*(1), 119–127.
4. Dascălu, M., Trăuşan-Matu, S., & Dessus, P. (2013). Cohesion-based analysis of CSCL conversations: Holistic and individual perspectives. In N. Rummel, M. Kapur, M. Nathan & S. Puntambekar (Eds.), *10th International Conference on CSCL 2013* (pp. 145–152). Madison, USA: International Society of the Learning Sciences.
5. Bakhtin, M. M. (1981). *The dialogic imagination: Four essays*. London: The University of Texas Press.
6. Trăuşan-Matu, Ş. (2010). The polyphonic model of hybrid and collaborative learning. In F. Wang, L. J. Fong & R. C. Kwan (Eds.), *Handbook of research on hybrid learning models: Advanced tools, technologies, and applications* (pp. 466–486). Hershey. NY: Information Science Publishing.
7. Murillo Montes de Oca, A., Nistor, N., Dascălu, M., & Trăuşan-Matu, Ş. (2014). Designing smart knowledge building communities. *International Journal of Interaction Design and Architecture(s), 22*, 9–21.